Am. J. Hum. Genet. 60:703–716, 1997

Detecting Disease-Predisposing Variants: The Haplotype Method

Ana M. Valdes and Glenys Thomson

Department of Integrative Biology, University of California at Berkeley, Berkeley

Summary

For many HLA-associated diseases, multiple alleles—and, in some cases, multiple loci—have been suggested as the causative agents. The haplotype method for identifying disease-predisposing amino acids in a genetic region is a stratification analysis. We show that, for each haplotype combination containing *all* the amino acid sites involved in the disease process, the relative frequencies of amino acid variants at sites not involved in disease but in linkage disequilibrium with the disease-predisposing sites are expected to be the *same* in patients and controls. The haplotype method is robust to mode of inheritance and penetrance of the disease and can be used to determine unequivocally whether all amino acid sites involved in the disease have *not* been identified. Using a resampling technique, we developed a statistical test that takes account of the nonindependence of the sites sampled. Further, when multiple sites in the genetic region are involved in disease, the test statistic gives a closer fit to the null expectation when *some*—compared with *none*—of the true predisposing factors are included in the haplotype analysis. Although the haplotype method cannot distinguish between very highly correlated sites in one population, ethnic comparisons may help identify the true predisposing factors. The haplotype method was applied to insulin-dependent diabetes mellitus (IDDM) HLA class II DQA1-DQB1 data from Caucasian, African, and Japanese populations. Our results indicate that the combination DQA1#52 (Arg predisposing) DQB1#57 (Asp protective), which has been proposed as an important IDDM agent, does not include all the predisposing elements. With rheumatoid arthritis HLA class II DRB1 data, the results were consistent with the shared-epitope hypothesis.

Received May 30, 1996; accepted for publication November 20, 1996.

Address for correspondence and reprints: Dr. Glenys Thomson, Department of Integrative Biology, 3060 Valley Life Sciences Building, University of California at Berkeley, Berkeley, CA 94720-3140. E-mail: glenys@violet.berkeley.edu

0002-9297/97/6003-0028$02.00

Introduction

The HLA-associated diseases are, in general, genetically complex (e.g., see Thomson 1988, 1995*a*). They can exhibit incomplete penetrance, the involvement of multiple HLA loci, genetic heterogeneity within the HLA region, synergistic effects, and the involvement of non-HLA loci. Difficulty in the identification of specific disease-predisposing and protective alleles at loci within the HLA region results from the fact that multiple genetic factors may be involved, including genetic variants that are common in the general population (e.g., see Tait and Harrison 1991; Thomson 1991). Also, amino acids—or a particular sequence of amino acids—involved in the disease process are difficult to identify in the context of the high linkage disequilibrium common within the HLA region, in particular the very strong disequilibrium of the HLA class II DR and DQ genes (Begovich et al. 1992; Imanishi et al. 1992).

Every HLA allele is defined by a unique DNA sequence in the exons. In the case of the HLA class II loci, most of the variation is confined to hypervariable regions in the second exon, which affects the antigen-binding pocket. The amino acids at variable sites in an allele typically occur in other alleles as well, which gives rise to the characteristic patchwork of variation seen when the amino acid composition of HLA alleles is compared (Lawlor et al. 1990). Such patterns of amino acid site variability raise the possibility that HLA-variation association with a disease may not be due to a given allele but, rather, to one or more variable amino acid sites (shared epitopes) occurring on several alleles. Shared epitopes are suggested to be responsible for the HLA class II DRB1 associations with rheumatoid arthritis (RA) (Gregersen et al. 1987). The hierarchy of relatively predisposing through protective effects of HLA alleles for insulin-dependent diabetes mellitus (IDDM) suggests that analysis at the amino acid—rather than at the allelic—level may be particularly informative (e.g., see Cucca et al. 1993, 1995; Yasunaga et al. 1996).

The original application of the haplotype method was to allele-frequency data (Thomson et al. 1988). Direct roles of HLA DR3 and DR4 in IDDM were excluded, since HLA B locus variation on these haplotypes was different in patients and controls. In this paper we present the theoretical basis of the haplotype method for testing whether all disease-predisposing variants have

been identified. The models are described with regard to amino acid variation, but the results apply equally to allelic variation. We consider separately the observed haplotypic amino acid combinations at putative predisposing sites. For each haplotype combination containing all the amino acid sites involved in disease process—for example, amino acids a_1 and b_1 at sites A and B—the relative frequencies of amino acid variants at sites not involved in disease—for example, r_1 and r_2 at site R—are expected to be the same in patients and controls. That is, although the absolute frequencies of the haplotypes $a_1b_1r_1$ and $a_1b_1r_2$ will differ between patients and controls, the relative frequency of the ratio $a_1b_1r_1/a_1b_1r_2$ will be the same in patients and controls. Inequality of this ratio is expected if all sites involved in the disease process—and in linkage disequilibrium with the sites under consideration—have *not* been identified. Using a resampling technique, we develop for the haplotype method a statistical test that takes account of the nonindependence of the amino acid sites sampled, to determine whether all amino acids involved in the disease process have been identified. We then study the effect of including some—but not all—of the amino acid sites involved in the disease in the haplotype analysis, to determine whether the large number of combinations of putative disease-predisposing amino acid sites can be reduced without loss of information. We apply the haplotype method to HLA-DQA1/-DQB1 variation in IDDM in three populations and to HLA-DRB1 variation in RA in a Norwegian population.

The Haplotype Method

The realization that models of disease are in fact equivalent to selection models and, hence, to hitchhiking models (Thomson 1977) simplifies the theoretical development of the haplotype method. An important advantage of "disease" models over the usual study of selection operating at the population level is that, in the study of disease processes, we observe the genetic composition of a population both before and after the selective process (disease) has operated. The general population (called "controls" throughout this paper) represents the before-selection genetic variability, the patient population that after selection. Further, continuing the analogy with selection models, we note that disease models investigate the genetic changes that take place in one generation of selection without recombination. In this context, we can ignore how haplotype frequencies evolve in patients and controls over time.

Amino acid site(s) involved in disease predisposition or protection are treated as the selected "locus" ("loci")—and nonpredisposing sites are treated as neutral "loci"—and are considered in the context of classic selection hitchhiking models. The effect that the predisposing site(s) have on nonpredisposing sites is examined. For an extensive treatment of the hitchhiking effect—that is, the effect of a selected locus on allele frequencies at linked neutral loci—the reader is referred to the work of Thomson (1977). It should be noted that the haplotype method described here assumes random mating in the general population and that patients and controls share the same ethnic background. The results obtained here may not hold for populations that deviate substantially from random mating or for data sets in which patients and controls are not ethnically matched.

To identify predisposing variants, we ask whether there is a pattern expected consistently for predisposing sites but not for nonpredisposing ones, or vice versa. We consider separately the observed haplotypic amino acid combinations at putative predisposing sites. For each of these haplotypes, we compare the ratios of polymorphic putative nonpredisposing sites in patients versus those in controls. Intuitively, if *all* predisposing sites are included in these haplotypes, then the variants at the neutral sites should be in the same ratio on a particular disease-predisposing haplotype, in patients and in controls. The frequency of these haplotypes will differ between patients and controls, but the relative frequency of neutral sites on a particular haplotypic combination of all predisposing sites will be the same in patients and controls. Conversely, if not all predisposing sites in linkage disequilibrium have been identified and considered in the haplotypes, then equality in the ratios, in patients and controls, of polymorphism at putative nonpredisposing sites on these haplotypes is not expected. We develop the analytical work to prove this, and we develop a statistical test to determine when the hypothesis that all predisposing sites have been identified can be accepted or rejected.

One-Predisposing-Site Model

Consider a three-site model with two amino acid residues at each site. At the first site, assumed to be predisposing, the amino acid residues are *A* and *a; A* is predisposing relative to *a*. Both residues at each of sites 2 (*B* and *b*) and 3 (*C* and *c*) are equivalent with regard to disease and are nonpredisposing. The following notation is used for residue frequencies:

$$p_A = \text{frequency of } A,\ q_A = 1 - p_A = \text{frequency of } a\ ;$$

$$p_B = \text{frequency of } B,\ q_B = 1 - p_B = \text{frequency of } b\ ;$$

$$p_C = \text{frequency of } C,\ q_C = 1 - p_C = \text{frequency of } c\ .$$

The frequencies of all eight haplotypes in the control population, denoted $x_1, x_2, \ldots, x_8$, can be written in terms of the three allele frequencies, three pairwise linkage disequilibrium terms, and one third-order disequilibrium term (e.g., see Robinson et al. 1991):

$$x_1 = f(ABC) = p_A p_B p_C + p_A D_{BC} + p_B D_{AC} + p_C D_{AB} + D_{ABC}\,; \quad (1a)$$

$$x_2 = f(ABc) = p_A p_B q_C - p_A D_{BC} - p_B D_{AC} + q_C D_{AB} - D_{ABC}\,; \quad (1b)$$

$$x_3 = f(AbC) = p_A q_B p_C - p_A D_{BC} + q_B D_{AC} - p_C D_{AB} - D_{ABC}\,; \quad (1c)$$

$$x_4 = f(Abc) = p_A q_B q_C + p_A D_{BC} - q_B D_{AC} - q_C D_{AB} + D_{ABC}\,; \quad (1d)$$

$$x_5 = f(aBC) = q_A p_B p_C + q_A D_{BC} - p_B D_{AC} - p_C D_{AB} - D_{ABC}\,; \quad (1e)$$

$$x_6 = f(aBc) = q_A p_B q_C - q_A D_{BC} + p_B D_{AC} - q_C D_{AB} + D_{ABC}\,; \quad (1f)$$

$$x_7 = f(abC) = q_A q_B p_C - q_A D_{BC} - q_B D_{AC} + p_C D_{AB} + D_{ABC}\,; \quad (1g)$$

$$x_8 = f(abc) = q_A q_B q_C + q_A D_{BC} + q_B D_{AC} + q_C D_{AB} - D_{ABC}\,. \quad (1h)$$

The relative penetrance values for the three genotypes at the disease predisposing locus are denoted as follows:

$$\begin{aligned} &\text{Relative penetrance for } AA = g_2 = s + w\,; \\ &\text{Relative penetrance for } Aa = g_1 = s + hw\,; \\ &\text{Relative penetrance for } aa = g_0 = s\,. \end{aligned} \quad (2a)$$

The second formulation is used later in the description of the disease models, with the restrictions $0 \leq s \leq 1$, $0 \leq h \leq 1$, $0 \leq w \leq 1$, where s is the frequency of sporadics, h is a mode of inheritance parameter, and w is the penetrance. The prevalence of the disease is denoted by T, with

$$T = g_2 p_A^2 + 2 g_1 p_A q_A + g_0 q_A^2\,. \quad (2b)$$

The expected haplotype deterministic frequencies among patients, denoted by $y_1, y_2, \ldots, y_8$, are derived as the haplotype frequencies in a population after selection without recombination, by use of the penetrance terms in equation (2a) as the selection parameters (Thomson 1977):

$$y_i = \frac{x_i(g_2 p_A + g_1 q_A)}{T} \quad i = \{1, 2, 3, 4\}\,; \quad (3a)$$

$$y_i = \frac{x_i(g_1 p_A + g_0 q_A)}{T} \quad i = \{5, 6, 7, 8\}\,. \quad (3b)$$

In this one-site-predisposing model, it is easy to see from equation (3a) that equality holds for the ratios in patients and controls of nonpredisposing-site variation (e.g., B and b) on haplotypes containing a variant at the predisposing site (A in this case):

$$\left[\frac{f(AB)}{f(Ab)}\right]_{\text{patients}} = \frac{(y_1 + y_2)}{(y_3 + y_4)} = \left[\frac{f(AB)}{f(Ab)}\right]_{\text{controls}} = \frac{(x_1 + x_2)}{(x_3 + x_4)}\,, \quad (4a)$$

and, from equation (3b),

$$\left[\frac{f(aB)}{f(ab)}\right]_{\text{patients}} = \frac{(y_5 + y_6)}{(y_7 + y_8)} = \left[\frac{f(aB)}{f(ab)}\right]_{\text{controls}} = \frac{(x_5 + x_6)}{(x_7 + x_8)}\,. \quad (4b)$$

Thus, as intuitively expected, for haplotypes containing a variant at the disease-predisposing amino acid site, A or a in this case, the relative frequency of "neutral" sites, B and b, is the same in patients and controls, even though the absolute frequencies of these haplotypes differ between patients and controls. These results can be written algebraically, in the following form:

$$\frac{(y_1 + y_2)(x_3 + x_4)}{(y_3 + y_4)(x_1 + x_2)} = 1.0\,; \quad (5a)$$

$$\frac{(y_5 + y_6)(x_7 + x_8)}{(y_7 + y_8)(x_5 + x_6)} = 1.0\,. \quad (5b)$$

Similar results hold if we substitute C (or c) for B (or b).

Note that the above results hold generally for any value of sporadics and any mode of inheritance. For our purposes, sporadic cases may be due to any kind of factor (genetic or environmental) not correlated (i.e., in linkage disequilibrium) with the amino acid site(s) under study. Thus, even if there are other important genetic predisposing elements for the disease being studied, as long as these are in linkage equilibrium with the genetic region being considered, such elements will not affect the ratio of the relative frequencies with respect to the predisposing site(s) in the genetic region being consid-

ered. This will usually be the case with loosely linked and unlinked genes.

On the other hand, we expect that the equality of the ratios in equations (4a) and (4b) will *not* hold if the predisposing site in this case has not been identified; that is,

$$\left[\frac{f(BC)}{f(Bc)}\right]_{\text{patients}} = \frac{(y_1 + y_5)}{(y_2 + y_6)} \neq \left[\frac{f(BC)}{f(Bc)}\right]_{\text{controls}} = \frac{(x_1 + x_5)}{(x_2 + x_6)}, \tag{6}$$

unless B and C are in linkage equilibrium with A. Let

$$V = \left[\frac{f(BC)}{f(Bc)}\right]_{\text{patients}} \cdot \left[\frac{f(Bc)}{f(BC)}\right]_{\text{controls}} = \frac{(y_1 + y_5)(x_2 + x_6)}{(y_2 + y_6)(x_1 + x_5)} \tag{7}$$

From equations (1) and (3) the ratio V equals 1.0 only if D_{ABC}, D_{AB}, and D_{AC} are all 0—that is, the disease-predisposing site is in complete linkage equilibrium with both neutral sites. For all other cases, no general rule applies to the value that V will take, since many parameters are involved. However, values close to 1 when the true predisposing sites are not included as the putative predisposing sites usually are seen only with a high frequency of sporadic cases of disease, for this genetic region. Thus, rejection of the equality of equations (4a) and (4b) indicates that the true predisposing factors have not been identified.

Results similar to those derived under a one-predisposing-site model apply for two-predisposing-sites models (see appendix A). The haplotype method applies equally well when two or more predisposing sites are involved and when more than one neutral site is associated with the site(s) under study. The ratio of frequencies of nonpredisposing polymorphic sites will be the same in patients and controls, on haplotypes with a particular amino acid combination when *all* predisposing sites are used. The ratio for the putative nonpredisposing polymorphic sites in patients and controls will differ if all predisposing sites are *not* included in the analysis. The only exception to this rule is if the nonpredisposing and additional predisposing sites are in linkage equilibrium with the predisposing sites under consideration.

Incomplete Combinations of Amino Acids

It is of interest to know, in addition, whether, with the haplotype approach, it is possible to distinguish not only having identified all the predisposing sites from not having identified all of them but having identified *some* predisposing sites from having identified *none* at all. With the four-site model described in appendix A—two predisposing sites, A (or a) and B(or b), and two nonpredisposing sites, C (or c) and D (or d)—the question can be rephrased as follows: Is the ratio of the quantity in patients over controls, $f(AD)/f(Ad)$, which includes both *one* predisposing site (of *two*) and a neutral polymorphic site, closer to 1.0 than is the ratio of the quantity in patients to controls, $f(CD)/f(Cd)$, which does *not* include a predisposing site? Note that replacing A with a, B, or b and C with c and using C (or c) instead of D (or d) in the first ratio are equivalent with respect to this question. Let

$$R = \left| 1 - \frac{\left[\frac{f(CD)}{f(Cd)}\right]_{\text{patients}}}{\left[\frac{f(CD)}{f(Cd)}\right]_{\text{controls}}} \right| - \left| 1 - \frac{\left[\frac{f(AD)}{f(Ad)}\right]_{\text{patients}}}{\left[\frac{f(AD)}{f(Ad)}\right]_{\text{controls}}} \right|. \tag{8}$$

If $R > 0$, the relative frequencies of nonpredisposing sites are more similar in patients and controls in the combination that includes one of two predisposing sites than these relative frequencies are in the combination that includes no predisposing sites at all. This quantity measures whether, given a set of residue frequencies and linkage-disequilibrium values, the relative frequencies are, in patients and controls, more similar when *one* of *two* predisposing sites is included than when *none* are included.

Even in the simplest case (fully recessive with no sporadics), the analytical expression for R (eq. [8]) is extremely complex (appendix A), particularly when we take into account that there are 32 constraints on the four-site linkage-disequilibrium values, which involve all four-site allele frequencies, six pairwise disequilibria, and four three-site disequilibria. Further, those 32 constraints yield new constraints on the third-order disequilibria (Robinson et al. 1991). All these constraints should be included when the possible values of R are derived.

Given the difficulty involved in deriving an analytical solution to this problem, we have chosen to estimate numerically how likely it is that $R > 0$ for actual HLA amino acid residue frequencies. We are not attempting to assess the statistical significance of a small positive value of R; rather, we want to know whether it is possible to reduce the number of combinations to be evaluated, by keeping those combinations of sites that give

Table 1

Properties of R (eq. [8])

	Frequency of $R > 0$ (%)	Average R (Range)
Caucasian:		
Recessive, $s = 0$	76.8	2,807.5 (−18.3, +6,085,229.2)
Dominant:		
$s = 0$	77.5	.162 (−1.007, +2.760)
$s = .25$	77.1	.082 (−.760, +.999)
Japanese:		
Recessive, $s = 0$	70.3	6.009 (−12.14, +3,645.4)
Dominant:		
$s = 0$	71.9	.131 (−5.63, +33.97)
$s = .25$	71.2	.065 (−1.843, +1.032)
African:		
Recessive, $s = 0$	80.1	5.814 (−15.86, +8,091.4)
Dominant:		
$s = 0$	80.7	.155 (−2.05, +6.97)
$s = .25$	79.4	.102 (−1.630, +1.062)

values more similar to what is expected if all predisposing sites are included.

We have calculated four-site, three-site, and two-site linkage-disequilibrium values for several combinations of four sites from DQA1-DQB1 haplotype-frequency data in Japanese, Caucasian, and African data sets from the 11th Histocompatibility Workshop (Rønningen et al. 1992). For simplicity, only sites that segregate for two residues were used. Using these data, we computed the quantity R above (eq. [8]) for 2,000 points chosen at random from each population. We used incomplete penetrance and allowed for nonzero values of s—that is, sporadic cases of the disease.

The mode of inheritance and the population studied made a substantial difference in terms of the *value* of R but were irrelevant in terms of its sign (table 1). The results were not affected even with moderate amounts of sporadic cases. Although $R > 0$ (eq. [8]) is *not* true *all* the time, it does hold in the majority (>70%) of cases (Valdes 1995).

Given the high polymorphism in the HLA region, the number of possible combinations of polymorphic sites to analyze with the haplotype method is very large unless the number of combinations to be considered can be reduced. The result from this section—namely, that the haplotype method more often will give a fit closer to the null expectation ratio of 1 (all factors have been identified) when *some,* compared with *none,* of the true factors are included in the haplotype analysis as putative predisposing sites—allows an initial screening of combinations of putative predisposing sites. Amino acid combinations that give the poorest fit can be removed from further consideration as putative predisposing sites in the search for fully predisposing combinations of sites. It must be noted that the method proposed is designed primarily as a hypothesis-testing method. By keeping those combinations that more closely resemble a fully predisposing set of sites (see the companion paper in this issue of the *Journal* [Valdes et al. 1997]), we are not following a strict mathematical algorithm. Other information also can be included in our decision of which sites to consider as putative predisposing combinations of amino acids—for example, the relative risks at each single polymorphic amino acid site, sites believed to be of functional importance in the structure of the HLA molecules, sites proposed after analysis by other methods such as the unique combinations method (Salamon et al. 1996), and sites that have been hypothesized by others as important in IDDM.

Computer Simulations

To apply these theoretical deterministic results to actual data, we must develop a statistical test capable of distinguishing predisposing sites from nonpredisposing sites in protein sequence data. The high linkage disequilibrium seen with amino acids of the HLA loci and other closely linked genetic regions must be taken into account. We generated protein sequences in a scenario where we know which site(s) have determined the differences between patients and controls. We devised two settings to do this—(1) a Fisher-Wright model where sequences are generated by use of a standard coalescent simulation and (2) actual HLA haplotype frequencies in a Norwegian data set (Rønningen et al. 1991). Once we have generated patient and control sequence data, we need to derive a statistic capable of testing the deterministic results of the haplotype method.

Assume first, for simplicity, that all amino acid sites that will be encountered segregate for only two residues and that we start looking at site i, which segregates for A and a. We then pick another polymorphic site, which segregates for B and b. If one of the residues at site i is predisposing and there are no other predisposing sites, equations (4a) and (4b) should hold. It is then possible to write a contingency table as depicted in table 2, with

Table 2

Contingency Table

	Patients	Controls	Total
$f(AB)$	w	x	$w + x$
$f(Ab)$	y	z	$y + z$
Total	$w + y$	$x + z$	N

NOTE.—$N = x + y + w + z$. The statistic $N(wz-xy)^2/[(w+x)(y+z)(w+y)(x+z)]$ follows a χ^2 distribution with 1 df under the null hypothesis that all predisposing sites have been identified.

the null hypothesis of homogeneity being that site i is the sole predisposing site.

If the null hypothesis is true, the statistic derived from the contingency table will follow a χ^2 distribution with 1 df. The analysis can be extended to compare the relative frequency of any number of residues segregating at site i and the residues at other polymorphic sites that segregate for more that two residues, by generating a $2 \times m$ (where m denotes the number of residues at the other site) contingency table for each residue at site i. In this case the test statistic should follow a χ^2 distribution with $m - 1$ df. It is important to discard from the analysis those contingency tables that have small sample sizes. For this reason we have included only those tables with a total sample size ≥ 20. We tested whether the measurement derived from the contingency table follows a χ^2 distribution, by plotting, in a quantile-quantile plot, simulation data against the expected χ^2 distribution data (Kendall and Stuart 1979).

Simulations Using Coalescent Data

We generated a gene genealogy of 100 sequences, using the simulation program of Hudson (1990), on the basis of the coalescent process for a neutral locus without recombination. As described by Hudson (1990), each sample is obtained by first producing a genealogy of the sample, under the assumption of a large constant population size. Once the genealogy is produced, mutations are randomly placed on the genealogy. At the root of the genealogy, a random ancestral amino acid sequence (100 sites long) is generated; mutation numbers corresponding to a Poisson distribution with parameters $2N_e\mu$ (N_e = effective population size, and μ = mutation rate per locus) and t (length of the branch), are generated with the genealogy. A predisposing site (i.e., the amino acid site at which the predisposing residue occurs) is defined arbitrarily. We defined the population frequency (q) of the predisposing amino acid, within limits—for example, $.3 < q < .4$—such that the program picked the first node that it found with 30–40 descendants. From this node on, all its descendants (unless a mutation occurred later) would carry a "P" at the predisposing site, whereas the ancestral sequence—and, therefore, all of the other descendants except those that mutate at that site—would carry an "N."

At the tips of the tree, we sampled pairs of sequences with replacement. Each of these pairs represents an individual. We then applied a disease model by defining the values of g_2, g_1, and g_0, from equation (2a). We kept sampling with replacement until we obtained a pool of 100 patient haplotypes and 100 control haplotypes.

Simulations Using HLA Data

In this case, instead of generating sequences, we used published haplotype frequency data for DRB1, DQA1, and DQB1 loci. From this control population, we generated ("selected") the patient population. We defined a site (or a combination of sites) as predisposing and picked at random the predisposing residue. A data set of DRB1-DQA1-DQB1 haplotypes for 181 unrelated Norwegian individuals (Rønningen et al. 1991) was used as the reference pool for selection of patients.

Quantile-Quantile Plots

In each simulation run, we built a contingency table of patients and controls, using as entries the frequencies of the two residues at a nonpredisposing site (segregating for only two residues) on haplotypes with a residue from the predisposing site, and derived the χ^2 value (see table 2). We repeated this process 1,000 times and graphed the data in a quantile-quantile plot. Both for the simulations that used sequences generated by a neutral coalescent genealogy and for those using HLA class II sequences, the test statistics computed by comparing polymorphic neutral sites on haplotypes with a residue from the predisposing site follow a χ^2 distribution, as revealed by the quantile-quantile plot (fig. 1, *upper panel*).

The exact same procedure was repeated, but now not using as a reference haplotypes with a residue from the predisposing site but using haplotypes with a residue from a nonpredisposing site, and another quantile-quantile plot was generated. The distribution of the test statistic when the predisposing site is not included is very different from the χ^2 (fig. 1, *lower panel*). Furthermore, the distributions of the test statistic for HLA data and for neutral coalescent sequences are also very different with respect to each other when the predisposing site is not included, indicating that the different nature of the sequence data (resulting from a different evolutionary history) affects the distribution of the test statistic, in nonpredisposing sites. Both distributions, however, are very different from the χ^2 distribution.

Inclusion of Data with >1 df

In general, we need to allow for sites segregating for more than two residues. The distribution of the number of residues at amino acid sites will depend on the nature of the data; for example, HLA sequence data segregate, on average, for more residues than do neutral coalescent data. However, this does not affect the general result: in computer simulations using as many as nine predisposing sites—and in which these sites segregate for as many as three residues—the test statistic computed with all predisposing sites included still followed a χ^2 distribution with the corresponding df (results not shown).

To allow for cases with varying df, which is due to different numbers of segregating sites, we propose a standardized χ^2 measure, defined as

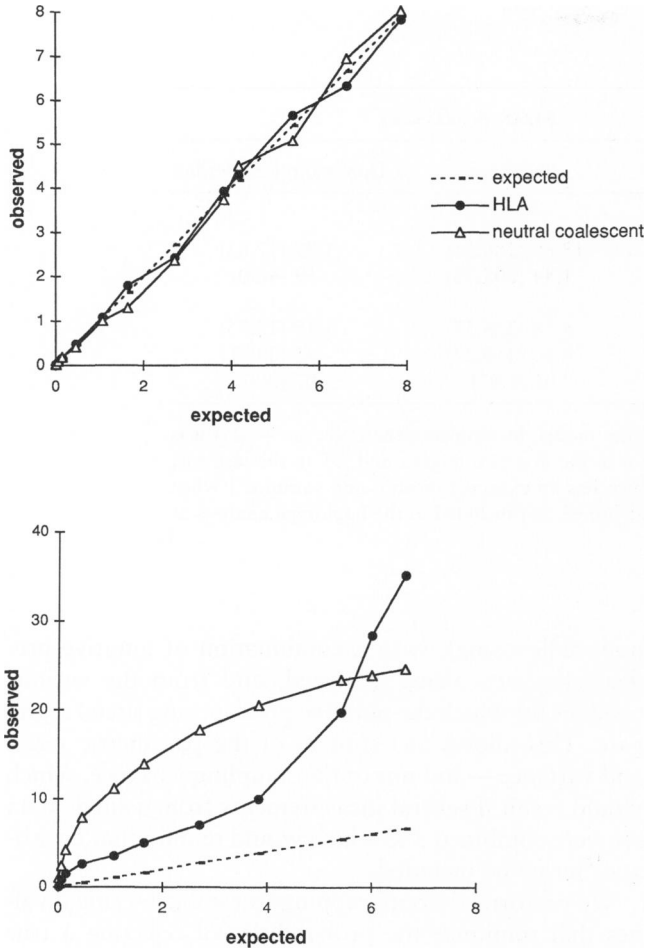

Figure 1 Quantile-quantile plots using HLA data and protein sequences generated by a coalescent simulation for a contingency-table test of homogeneity when (*upper panel*) the true predisposing site is included in the haplotype analysis and (*lower panel*) a nonpredisposing site is used as the putative predisposing site. The expected values correspond to a χ^2 distribution with 1 df.

$$\overline{X^2} = \frac{\chi^2 - \nu}{\sqrt{2\nu}}, \qquad (9)$$

where χ^2 is the observed statistic from the contingency-table test of homogeneity and where ν is the number of df in the contingency table. The standardized χ^2 measure for haplotypes including *all* predisposing sites will then have a mean of 0.0 and a variance of 1.0.

We computed the χ^2 measure (table 2) and standardized χ^2 measure (eq. [9]) test statistics, using, at first, only contingency tables with 1 df, by choosing neutral (nonpredisposing) sites that segregated for only two residues. As expected, χ^2 values with 1 df for the true predisposing site give an estimated mean very close to 1.0 and an estimated variance of ~2.0, for both types of simulation (table 3A), and give standardized χ^2s with a mean close to 0 and variance of ~1.0 (table 3B).

Measurements corresponding to haplotypes with nonpredisposing sites considered as putative predisposing sites depart substantially from both the expected mean and variance values, when the true predisposing sites are used (tables 3A and 3B). The mean χ^2 values with 1 df obtained for nonpredisposing sites would be exceeded with a probability of ~.15. Large differences between coalescent and HLA nonpredisposing sites also were observed. When simulations using contingency tables with >1 df were run, the observed values of the mean and variance of the standardized χ^2 test statistic (eq. [9]) were, again, very close to the expected values for predisposing sites, which are mean 0 and variance 1, but not to those for nonpredisposing sites (table 3C).

We also considered, for HLA data, cases where there are two predisposing sites and two different modes of inheritance: dominant and recessive and also sporadic cases (table 4). The results in all of these situations are very similar when we base the analysis on the pair of predisposing sites. Results for nonpredisposing sites are, in all cases, very different from those for predisposing sites, but they vary substantially with mode of inheritance.

Previously we noted that, if, in the analysis, we include as putative predisposing sites *some*—but *not all*—of the

Table 3

Computer-Simulation Results: One Predisposing Site

Statistic and Site[a]	Mean (Variance)[b]
A. χ^2 with 1 df:	
Coalescent:	
Predisposing	1.006 (2.172)
Nonpredisposing	2.466 (12.631)
HLA:	
Predisposing	1.000 (1.979)
Nonpredisposing	2.385 (27.634)
B. Standardized χ^2 with 1 df:	
Coalescent:	
Predisposing	.004 (1.086)
Nonpredisposing	.869 (5.637)
HLA:	
Predisposing	.000 (.988)
Nonpredisposing	1.432 (13.493)
C. Standardized χ^2 with mixed df:	
Coalescent:	
Predisposing	−.044 (.970)
Nonpredisposing	.824 (5.209)
HLA:	
Predisposing	−.069 (.932)
Nonpredisposing	2.059 (17.460)

[a] "Predisposing" denotes a putative predisposing site that is analyzed by the haplotype method and that is the true predisposing site; and "nonpredisposing" denotes that the true predisposing factor is not being considered.

[b] Data are from computer simulations using one predisposing site and a dominant mode of inheritance with incomplete penetrance.

Table 4

HLA Simulation Results: Two Predisposing Sites

	MEAN (VARIANCE)		
	Dominant[a]	Recessive	Dominant + Sporadics
Single Site:			
Nonpredisposing site	2.06 (23.37)	13.89 (286.51)	1.89 (17.65)
One predisposing site	1.44 (14.15)	1.44 (205.73)	.59 (4.80)
Analysis by pairs:			
Nonpredisposing/nonpredisposing	1.63 (23.12)	9.58 (226.55)	1.19 (15.75)
Predisposing/nonpredisposing	1.07 (16.73)	6.10 (150.22)	.35 (3.07)
Predisposing/predisposing	−.09 (.89)	−.01 (1.01)	−.07 (.88)

[a] Results from simulations under a two-predisposing-sites model. Incomplete penetrance (w = .25) was used in all cases. The penetrance of sporadic cases was 0 in the first two models and .05 in the last one. Values are for the standardized χ^2 measure (eq. [9]), which has an expected mean 0 and variance 1 when all (i.e., both in this case) predisposing sites have been identified and included in the haplotype analysis as the putative predisposing sites.

predisposing sites, the result, in terms of the test statistic, in most cases should be more similar to having *all* predisposing sites than to having *no* predisposing sites at all. This is confirmed by the results displayed in table 4, either when sites are analyzed by pairs (including one predisposing site and one nonpredisposing, as opposed to including two nonpredisposing sites) or when they are analyzed alone (including one of the predisposing sites vs. including one single nonpredisposing site). The addition of sporadic cases does not affect the estimated values of the mean and variance for the predisposing pair, although the values for the test statistic are lower with nonpredisposing sites used as putative predisposing sites than in the simulations where sporadic cases were not included. This suggests that, as sporadic effects become more important relative to the genetic factor under consideration, the ratio of the relative frequencies for nonpredisposing cases would be more similar in patients versus controls. As intuitively expected, it would then be more difficult to detect the actual predisposing sites.

Resampling: Type I and Type II Errors

The simulation results discussed so far involve independent replicates; they do not address the lack of independence that arises (1) from combining measurements corresponding to the various residues present at the predisposing site(s) and (2) from all the putative nonpredisposing sites considered separately with the putative predisposing site(s). In practice, to compute a mean and variance of the standardized χ^2 test statistic (eq. [9]), we must combine several nonindependent measurements. We propose the use of a resampling technique (bootstrapping) (Efron 1982), to resolve this issue (see below). This approach allows us to incorporate information from more than one amino acid site (presumed to be nonpredisposing), with a combination of putative predisposing sites being analyzed, and from the various residues for which the putative predisposing site(s) segregate. This allows an estimate of the parametric mean and variance—and not of the sampling variance, which would result if several measurements from a single data set were combined and which would require that covariance terms be included.

We performed bootstrapping to establish critical values that minimize the probabilities of rejecting a true null hypothesis (type I error) and of accepting a false null hypothesis (type II error). The resampling procedure consists of taking subsamples of 50 patients and 50 controls (100 haplotypes each) from a larger data set. Each time, a standardized χ^2 test statistic (eq. [9]) is computed for a residue (haplotype) from the combination of sites (or a single site) presumed to include all the predisposing sites and from a reference putative nonpredisposing site chosen randomly. We repeat this process 50–100 times and, at the end, estimate the parametric mean and variance.

We derived the joint distribution of the parameters of interest—namely, the mean and variance of the standardized χ^2. For complete predisposing combinations, the intervals $-0.2 <$ mean < 0.2 and $0.5 <$ variance < 1.8 include 97% of all mean and variance pairs (fig. 2). It should be noted that these intervals have been derived ad hoc for HLA class II data for a pair of predisposing sites, so the values presented can be applied to any of the many autoimmune and infectious diseases associated with HLA class II. For other types of sequence data—for example, neutral coalescent data—these critical values may vary and must be calculated for each type of data.

To assess type I and type II errors, we chose four

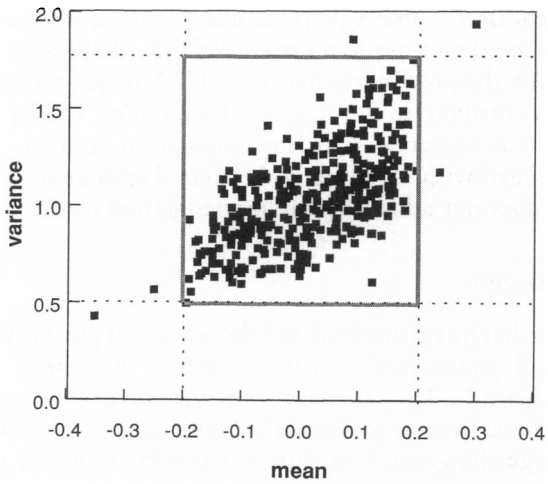

Figure 2 Joint distribution, from computer simulations, for HLA class II data on the mean and variance of the standardized χ^2 (eq. [9]) when all predisposing sites are included in the haplotype analysis derived from 20,000 runs (resampled into 500 groups of 50, from which 500 mean and variance values were computed).

pairs of sites from DQA1-DQB1. Each of these sites segregates for only two residues in a large data set of Caucasians (Rønningen et al. 1991) from which we took the DQA1-DQB1 haplotype frequencies for controls. For each of the four pairs of sites, which have varying amounts of linkage disequilibrium (table 5), we generated imaginary patient populations, assuming that the reference residues were predisposing; the disease models used are detailed in appendix B. We subsampled the 24 fictitious patient data sets (6 for each pair of sites, each set consisting of 1,000 DQA1-DQB1 haplotypes) 5,000 times each. A standardized χ^2 test statistic was computed each time, and 100 subsets of 50 individuals were formed. For each of these subsets the mean and variance were computed, and the number of times that $-0.2 < \text{mean} < 0.2$ and $0.50 < \text{variance} < 1.8$ was determined.

Type I Error

The average type I error observed (i.e., percent of times that the mean and variance fell outside the fixed intervals) was 1.78% ± 1.14%, averaged over the four pairs of sites and the six disease models. This value is consistent with the 3% error obtained when critical values were derived (fig. 2). There was no effect due to either the disease model or the amount of linkage disequilibrium between the sites involved in the disease.

Type II Error

It is not possible to analyze all the possible combinations to assess the actual type II error under all the situations that we can encounter with HLA data, because of the very large number of possible combinations, so we considered two pairs of sites: DQB1#66-DQB1#75 (D_{ij} = −0.139) and DQB1#38-DQB1#77 (D_{ij} = −0.243) and have used 10 other pairs that, when taken together with these "predisposing" sites, have varying levels of linkage disequilibrium. We then computed the percent of times (in the 100 subsets described previously) that the nonpredisposing pairs fell *within* the specified intervals, under the null hypothesis that all predisposing sites have been identified, and we used this number as our estimate for type II error. When the frequency of sporadics was 0 or only moderate, type II error was 0% for all disease models and all pairs. In cases with very high sporadic rates, type II error ranged from 8.5% (±17.02%) to 81.6% (±10.33%). This is not unexpected, since, in this case, there is very little distinction between the genetic and sporadic cases of disease, and since most cases are sporadics.

Application of the Haplotype Method to Autoimmune Diseases

IDDM

IDDM is one of the most studied HLA-associated diseases (Thomson 1988, 1995*a*). The pattern of inheritance of IDDM is complex, and there is no ubiquitous HLA haplotype associated with the disease. Although multiple non-HLA genes have been related to the disease, the HLA-linked genes are the major susceptibility markers known to date (e.g., see Todd 1995).

The HLA-DQ association with IDDM has been explained by a functional hypothesis in which DQB alleles encoding aspartic acid (Asp) in the 57th codon (Asp57) protect against IDDM, whereas non-Asp57 alleles confer susceptibility. Both predisposing and susceptible DQB1 alleles not fitting the position 57 DQ rule have been reported in several populations (Todd et al. 1989; Jenkins et al. 1990; Vicario et al. 1992). Because susceptibility to IDDM also correlates with DQA1 variation, the one-amino-acid model for susceptibility to IDDM was extended to a more sophisticated model with non-Asp57 DQβ chains and arginine 52 DQα chains predisposing in DQαβ dimers (Khalil et al. 1990). This simple model has been supported but also challenged in many ways (e.g., see Tait and Harrison 1991; Robinson et al.

Table 5

Pairs of Sites Used for Estimating Type I and Type II Errors

i, j	Reference Residues	D_{ij}
DQB1 66, DQB1 75	E, V	−.1387
DQB1 38, DQB1 77	V, R	.2426
DQA1 40, DQB1 38	E, V	−.0070
DQA1 34, DQB1 56	E, P	.0055

Table 6

Application of Haplotype Method to DQA1 #52 and DQB1 #57 in IDDM

	Mean (Variance)[a]
Caucasian	.8161 (3.0408)
African	.7655 (4.3614)
Japanese	.8133 (4.7877)
Expected range	−.2, .2 (.5, 1.8)

[a] Class II data from the 11th HLA workshop (Rønningen et al. 1992).

1993): transgenic mouse experiments have been confirmatory (Acha-Orbea and McDevitt 1987).

We applied the haplotype-resampling method described above to three different populations from the 11th Histocompatibility Workshop (Rønningen et al. 1992): Caucasian, African, and Japanese. For all three ethnic groups, the mean and variance from resampling of the data, with DQB1#57 and DQA1#52 used as the two predisposing sites, fell *outside* the expected range derived for HLA class II data (table 6). This means that DQA1#52 and DQB1#57 are *not* enough to account for the HLA predisposition to IDDM in the three ethnic groups analyzed. Further research is required, to establish whether any combination of sites within the DR-DQ class II loci can explain the HLA component for IDDM susceptibility (see the companion paper [Valdes et al. 1997]).

RA

RA is associated with HLA DRB1 alleles: in Caucasians, the frequencies of alleles DRB1*0401 and DRB1*0404 are increased in patients compared with controls (Nepom et al. 1986); the frequency of DRB1*0405 is increased in Japanese patients (Watanabe et al. 1989). An excess of DRB1*0101 also has been reported (Shiff et al. 1982). At residues 67–74, the amino acid sequence of DRB1*0404 and DRB1*0405 is identical, whereas the sequence of DRB1*0401 differs at only one site; and, at these same sites, the chain encoded by DRB1*0101 is identical to that encoded by DRB1*0404 and DRB1*0405. It therefore has been proposed that RA may be associated primarily with a shared epitope involving residues 67–74 (Gregersen et al. 1987).

We have tested this hypothesis by applying the haplotype-resampling method to a data set of DRB1 allele frequencies in 54 Caucasian adults diagnosed with RA and in 181 healthy controls (Rønningen et al. 1990). We resampled the data 100 times, with sites 67, 70, 71, and 74 considered as the putative predisposing sites. The mean and variance of the standardized χ^2 measure from this data were 0.15 and 1.01, respectively. These values fall within the intervals previously defined, supporting the shared-epitope hypothesis for RA. This does not exclude the possibility that other HLA-region variation may contribute to RA (e.g., see Dizier et al. 1993); but, if that is so, at least in the population studied, it is correlated with variation at the shared epitope in a way that does not allow the haplotype method to detect it.

Discussion

The haplotype method and the statistical test for HLA class II amino acids, as developed in this paper, test whether all relevant amino acid sites involved in a disease have been identified. They may prove valuable in distinguishing between different combinations of sites that have been proposed as determinants for genetic diseases and, if the study is applied to diverse ethnic groups, in choosing appropriate combinations of sites for risk assessment. Extensions of this method to genotypic effects will further increase its usefulness (the authors' work in this area is in progress).

We have shown that amino acid variation at sites DQA1#52 and DQB1#57 is not enough to explain the major HLA association with IDDM. This has been pointed out by others, using different approaches (e.g., see Tait and Harrison 1991; Robinson et al. 1993), but we use this example as demonstration of the ability of the haplotype method to detect that not all predisposing sites have been identified. This result clearly points to the importance of finding other combinations that may explain the class II component of IDDM in a more satisfactory way. In the companion paper (Valdes et al. 1997), a large number of possible combinations of sites within the DRB1-DQA1-DQB1 loci in different ethnic groups are examined. On the other hand, at least in the Caucasian data set analyzed, the class II DRB1 shared-epitope hypothesis cannot be excluded as the sole HLA component of susceptibility to RA.

It is important to note that any combination of sites that fits the use of the haplotype-method criteria is valuable only for its predictive value in terms of disease susceptibility. We are unable to make a statement regarding the functional involvement, in the disease process, of the molecular variants under study. Thus, there may be several combinations of sites, correlated with each other, capable of *predicting* disease susceptibility/protection in a certain genetic region. In order to determine the functional significance that various sites may have, molecular modeling and mutation analysis are needed.

The method that herein has been developed is designed for hypothesis testing and not as an algorithm for detecting amino acid combinations more likely to be involved in disease susceptibility. Although, both in this paper and in the companion paper (Valdes et al. 1997),

we have looked at diseases for which a hypothesis already is available, for other, less-studied diseases, direct application of the haplotype method might not be the best starting point. As already has been mentioned, one possibility is to compare the relative risks at each single polymorphic site or at each pair of sites, before the haplotype method is applied. A more sophisticated and accurate alternative is provided by the unique-combinations method developed by Salamon et al. (1996). This method detects all amino-acid-site combinations that distinguish a particular sequence—or set of sequences—from another set of sequences (e.g., patient and control populations).

The haplotype method can be applied to complex diseases in general and, in fact, to all diseases. However, it is obviously most relevant to those disorders (or traits) in which the alleles or amino acid involved in a disease are reasonably frequent. In the case of fully penetrant, monogenic diseases with predisposing alleles that are relatively rare in the population—for example, cystic fibrosis—the haplotype method may be difficult to apply, since appropriate control (population) data on haplotypes with the disease-predisposing variants are not normally available. With complex diseases, nuclear family–based data also can be used, as well as case-control data. With simplex families ascertained for the presence of an affected child, the parental alleles (haplotypes) not transmitted to the affected child form the AFBAC (affected-family–based control) population and give an unbiased estimate of population control-allele frequencies in the case of a random mating population and zero recombination between the marker and disease (see Thomson 1995*b*).

The major advantage of the haplotype method is that it allows hypothesis testing. Minimally, it can play a major role in determining when *not all* the sites involved in a disease in a genetic region have been identified. Further, it can indicate which sites are *more* likely involved than others. We are confident that application of this method will help us gain insights into the role that different loci and amino acids may play in susceptibility to complex autoimmune diseases.

Acknowledgments

We thank Shannon McWeeney and Terry Speed for helpful discussions and comments on the manuscript. This work was supported by NIH grants HD12731 and GM35326.

Appendix A

Two Predisposing-Sites Models

We extend the model to two predisposing sites and two nonpredisposing sites, each segregating for two amino acid residues. Let x_i, $i = 1, 2, \ldots, 16$, denote the frequencies of the four-locus haplotypes among control individuals:

$$x_1 = f(ABCD); \quad x_2 = f(ABCd);$$
$$x_3 = f(ABcD); \quad x_4 = f(ABcd);$$
$$x_5 = f(AbCD); \quad x_6 = f(AbCd);$$
$$x_7 = f(AbcD); \quad x_8 = f(Abcd);$$
$$x_9 = f(aBCD); \quad x_{10} = f(aBCd);$$
$$x_{11} = f(aBcD); \quad x_{12} = f(aBcd);$$
$$x_{13} = f(abCD); \quad x_{14} = f(abCd);$$
$$x_{15} = f(abcD); \quad x_{16} = f(abcd).$$

The definition of all 16 four-locus haplotypes in terms of allele frequencies and of 2d-, 3d-, and 4th-order linkage-disequilibrium terms can be found elsewhere (e.g., see Robinson et al. 1991).

We assume that residue *A* at the first site is predisposing and *a* is not; that residue *B* at the second site is predisposing and *b* is nonpredisposing; and that *C*, *c*, *D*, and *d* are all nonpredisposing at sites 3 (*C*/*c*) and 4 (*D*/*d*). A matrix is used to define the disease-penetrance parameters in the general case,

	AA	*Aa*	*aa*
BB	g_{22}	g_{12}	g_{02}
Bb	g_{21}	g_{11}	g_{01}
bb	g_{20}	g_{10}	g_{00}

In terms of the parameters as in equation (2a), the penetrance terms for disease models also can be expressed as

	AA	*Aa*	*aa*
BB	$s + w$	$s + hw$	s
Bb	$s + hw$	$s + h^2w$	s
bb	s	s	s

The case $s = 0$ corresponds to the existence of no sporadics. For recessive models $h = 0$, and for dominant models $h = 1$.

Let

$$U = g_{22}f(AB) + g_{21}f(Ab) + g_{12}f(aB) + g_{11}f(ab);$$
$$V = g_{21}f(AB) + g_{20}f(Ab) + g_{11}f(aB) + g_{10}f(ab);$$
$$W = g_{12}f(AB) + g_{11}f(Ab) + g_{02}f(aB) + g_{01}f(ab);$$
$$Z = g_{11}f(AB) + g_{10}f(Ab) + g_{01}f(aB) + g_{00}f(ab);$$
$$T = Uf(AB) + Vf(Ab) + Wf(aB) + Zf(ab).$$

The haplotype frequencies among patients are now given by

$$y_i = \frac{x_i U}{T} \quad i = \{1, 2, 3, 4\}\ ; \tag{A2}$$

$$y_i = \frac{x_i V}{T} \quad i = \{5, 6, 7, 8\}\ ; \tag{A3}$$

$$y_i = \frac{x_i W}{T} \quad i = \{9, 10, 11, 12\}\ ; \tag{A4}$$

$$y_i = \frac{x_i Z}{T} \quad i = \{13, 14, 15, 16\}\ . \tag{A5}$$

As in the one-predisposing-site model, the relative frequencies of polymorphic sites not involved in disease predisposition are the same in patient and control groups, on haplotypes including all disease-predisposing sites; that is, from equations (A2)–(A5),

$$\left[\frac{f(ABD)}{f(ABd)}\right]_{\text{patients}} = \left[\frac{f(ABD)}{f(ABd)}\right]_{\text{controls}};$$

$$\left[\frac{f(AbD)}{f(Abd)}\right]_{\text{patients}} = \left[\frac{f(AbD)}{f(Abd)}\right]_{\text{controls}};$$

$$\left[\frac{f(aBD)}{f(aBd)}\right]_{\text{patients}} = \left[\frac{f(aBD)}{f(aBd)}\right]_{\text{controls}};$$

$$\left[\frac{f(abD)}{f(abd)}\right]_{\text{patients}} = \left[\frac{f(abD)}{f(abd)}\right]_{\text{controls}}.$$

The same is true if, in the above equations, instead of using D and d, we use C and c. These results are equivalent to

$$\frac{(y_1 + y_3)(x_2 + x_4)}{(y_2 + y_4)(x_1 + x_3)} = 1\ ;$$

$$\frac{(y_5 + y_7)(x_6 + x_8)}{(y_6 + y_8)(x_5 + x_7)} = 1\ ;$$

$$\frac{(y_9 + y_{11})(x_{10} + x_{12})}{(y_{10} + y_{12})(x_9 + x_{11})} = 1\ ;$$

$$\frac{(y_{13} + y_{15})(x_{14} + x_{16})}{(y_{14} + y_{16})(x_{13} + x_{15})} = 1\ .$$

Equality also holds if additional neutral sites are incorporated into the haplotypes that include all predisposing sites; for example,

$$\left[\frac{f(ABCD)}{f(ABCd)}\right]_{\text{patients}} = \left[\frac{f(ABCD)}{f(ABCd)}\right]_{\text{controls}},$$

since $(y_1x_2)/y_2x_1) = 1$, from equation (A2)—and similarly for other combinations of $A(a)$, $B(b)$, and $C(c)$ or if $C(c)$ and $D(d)$ are exchanged. Haplotype combinations that do not include both predisposing sites $A(a)$ and $B(b)$, when the relative frequencies of neutral sites are considered, do not show this equality of appropriate ratios between patients and controls; for example,

$$\left[\frac{f(AD)}{f(Ad)}\right]_{\text{patients}} = \frac{(y_1 + y_3 + y_5 + y_7)}{(y_2 + y_4 + y_6 + y_8)} \neq \left[\frac{f(AD)}{f(Ad)}\right]_{\text{controls}} = \frac{(x_1 + x_3 + x_5 + x_7)}{(x_2 + x_4 + x_6 + x_8)}.$$

Equality in this case will occur only when the neutral site D (d) or C (c) is in complete linkage equilibrium with both predisposing sites and when the predisposing sites are in linkage equilibrium with each other.

Analytical Expression for R

The analytical expression of R (eq. [8]) when the mode of inheritance is fully recessive and when the penetrance of sporadic cases is 0 is given by

$$R = \left|1 - \frac{Z_1[p_C(1 - p_D) - D_{CD}]}{p_C p_D + D_{CD}}\right| - \left|1 - \frac{Z_2[p_A(1 - p_D) - D_{AD}]}{p_A p_D + D_{AD}}\right|;$$

$$Z_1 = \frac{Z_3(p_A p_B - D_{AB}) + Z_4}{q_D(p_C D_{AB} + p_B D_{AC} + p_A p_B p_C + p_A D_{BC} + D_{ABC}) - Z_5};$$

$$Z_2 = \left[1 + \frac{p_A p_B - p_A p_B p_D + D_{AB}}{-D_{ABD} - p_A D_{BD} - p_B D_{AD} - p_D D_{AB} - p_A p_B p_D}\right]^{-1};$$

$$Z_3 = p_A q_B p_C p_D - D_{ABCD} - p_C p_D D_{AB} + p_B p_D D_{AC} + q_B p_C D_{AD} - p_A p_D D_{BC} - D_{AD} D_{BC} - p_A p_C D_{BD} - D_{AC} D_{BD} + p_A q_B D_{CD} - D_{AB} D_{CD} - p_D D_{ABC} - p_C D_{ABD} + q_B D_{ACD} - p_A D_{BCD}\ ;$$

$$Z_4 = D_{ABCD} + p_A p_B p_C p_D + p_C p_D D_{AB} + p_B p_D D_{AC} + p_B p_C D_{AD} + p_A p_D D_{BC} + D_{AD} D_{BC} + p_A p_C D_{BD} + D_{AC} D_{BD} + p_A p_B D_{CD} + D_{AB} D_{CD} + p_D D_{ABC} + p_C D_{ABD} + p_B D_{ACD} - p_A D_{BCD}\ ;$$

$$Z_5 = D_{ABCD} + D_{AD}(D_{BC} + p_B p_C) + D_{BD}(D_{AC} + p_A p_C) + D_{CD}(D_{AB} + p_A p_B) + p_C D_{ACD} + p_A D_{BCD} .$$

Appendix B

Disease Models for Testing Type I and Type II Errors

In the models considered, the penetrance values of equation (A1) are as follows:

Dominant, no sporadic cases

$$\begin{matrix} .5 & .5 & 0 \\ .5 & .5 & 0 \\ 0 & 0 & 0 \end{matrix}$$

Recessive, no sporadic cases

$$\begin{matrix} .5 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{matrix}$$

Dominant, moderate penetrance of sporadic cases

$$\begin{matrix} .5 & .5 & .075 \\ .5 & .5 & .075 \\ .075 & .075 & .075 \end{matrix}$$

Recessive, moderate penetrance of sporadic cases

$$\begin{matrix} .5 & .075 & .075 \\ .075 & .075 & .075 \\ .075 & .075 & .075 \end{matrix}$$

Dominant, high penetrance of sporadic cases

$$\begin{matrix} .5 & .5 & .25 \\ .5 & .5 & .25 \\ .5 & .5 & .25 \end{matrix}$$

Recessive, high penetrance of sporadic cases

$$\begin{matrix} .5 & .25 & .25 \\ .25 & .25 & .25 \\ .25 & .25 & .25 \end{matrix}$$

References

Acha-Orbea H, McDevitt HO (1987) The first external domain of the nonobese diabetic mouse class II I-A b is unique. Proc Natl Acad Sci USA 84:2435–2439

Begovich AB, McClure GR, Suraj VC, Helmuth RC, Fildes N, Bugawan TL, Erlich HA, et al (1992) Polymorphism, recombination, and linkage disequilibrium within the HLA class II region. J Immunol 148:249–258

Cucca F, Lampis R, Frau F, Masis D, Anguis E, Masile P, Chessa M, et al (1995) The distribution of DR4 haplotypes in Sardinia suggests a primary association of type I diabetes with DRB1 and DQB1 loci. Hum Immunol 43:301–308

Cucca F, Muntoni F, Lampis R, Frau F, Argiolas J, Silvetti M, Anguish E, et al (1993) Combinations of specific DRB1, DQA1, DQB1 haplotypes are associated with insulin-dependent diabetes mellitus in Sardinia. Hum Immunol 37:85–94

Dizier M-H, Eliaou J-F, Babron M-C, Combe B, Sany J, Clot J, Clerget-Darpoux F (1993) Investigation of the HLA component involved in rheumatoid arthritis (RA) by using the marker association–segregation χ^2 (MASC) method: rejection of the unifying-shared-epitope hypothesis. Am J Hum Genet 53:715–721

Efron B (1982) The jackknife, the bootstrap, and other resampling plans. Society for Industrial and Applied Mathematics, Philadelphia

Gregersen PK, Silver J, Winchester RJ (1987) The shared epitope hypothesis: an approach to understanding the molecular genetics of susceptibility to rheumatoid arthritis. Arthritis Rheum 30:1205–1213

Hudson RR (1990) Gene genealogies and the coalescent process. Oxf Surv Evol Biol 7:1–44

Imanishi T, Azaka T, Kimura A, Tokunaga K, Gojobori T (1992) Allele and haplotype frequencies for HLA and complement loci in various ethnic groups. In: Tsuji K, Aizawa M, Sasazuki T (eds) Proceedings of the 11th International Histocompatibility Workshop and Conference. Vol 1: HLA 1991. Oxford University Press, Oxford, pp 1065–1220

Jenkins D, Mijovic C, Fletcher J, Jacobs KH, Bradwell AR, Barnett AH (1990) Identification of susceptibility loci for type I (insulin dependent) diabetes by trans racial gene mapping. Diabetologia 33:387–395

Kendall M, Stuart A (1979) The advanced theory of statistics. Macmillan, New York

Khalil I, d'Auriol L, Gobet M, Morin L, Lepage V, Deschamps I, Park MS, et al (1990) A combination of HLA-DQβ Asp57-negative and HLA DQα Arg52 confers susceptibility to insulin-dependent diabetes mellitus. J Clin Invest 85:1315–1319

Lawlor DA, Zemmour J, Ennis PD, Parham P (1990) Evolution of class I MHC genes and proteins: from natural selection to thymic selection. Annu Rev Immunol 8:23–63

Nepom GT, Hansen J, Nepom B (1986) The molecular basis for HLA class II association with rheumatoid arthritis. J Clin Immunol 7:1–7

Robinson WP, Asmussen MA, Thomson G (1991) Three-locus systems impose additional constraints on pairwise disequilibria. Genetics 129:925–930

Robinson WP, Thomson G, Barbosa J, Rich SS (1993) The homozygous parents affected sib pair method of detecting disease predisposition effects: application to insulin dependent diabetes mellitus. Genet Epidemiol 10:273–288

Rønningen KS, Spurkland A, Egeland T, Iwe T, Munthe E, Vartdal F, Thorsby E (1990) Rheumatoid arthritis may be primarily associated with HLA-DR4 molecules sharing a particular sequence at residues 67–74. Tissue Antigens 36: 235–240

Rønningen KS, Spurkland A, Iwe T, Vartdal F, Thorsby E (1991) Distribution of HLA-DRB1, -DQA1 and -DQB1 alleles and DQA1-DQB1 genotypes among Norwegian patients with insulin-dependent diabetes mellitus. Tissue Antigens 37:105–111

Rønningen KS, Spurkland A, Tait BD, Drummond B, Lopez-Larrea C, Baranda FS, Menendez-Diaz MJ, et al (1992) HLA class II associations in insulin-dependent diabetes mellitus among Blacks, Caucasoids, and Japanese. In: Tsuji K, Ai-

zawa M, Sasazuki T (eds) Proceedings of the 11th International Histocompatibility Workshop and Conference. Vol 1: HLA 1991. Oxford University Press, Oxford, pp 713–722

Salamon H, Tarhio J, Rønningen KS, Thomson G (1996) On distinguishing unique combinations in biological sequences. J Comp Biol 3:407–423

Shiff B, Mizrachi Y, Orgad S, Yaron M, Gazit E (1982) Association of HLA-Aw31 and HLA-DR1 with adult rheumatoid arthritis. Ann Rheum Dis 41:403–404

Tait BD, Harrison LC (1991) Overview: the major histocompatibility complex and insulin dependent diabetes mellitus. In: Harrison LC, Tait BD (eds) The genetics of diabetes. Part 1. Bailliere Tindall, London, pp 211–228

Thomson G (1977) The effect of a selected locus on linked neutral loci. Genetics 85:753–788

——— (1988) HLA disease associations: models for insulin dependent diabetes mellitus and the study of complex human genetic disorders. Annu Rev Genet 22:31–50

——— (1991) HLA population genetics. In: Harrison LC, Tait BD (eds) The genetics of diabetes. Part 1. Bailliere Tindall, London, pp 247–260

——— (1995*a*) HLA disease associations: models for the study of complex human genetic disorders. Crit Rev Clin Lab Sci 32:183–219

——— (1995*b*) Mapping disease genes: family-based association studies. Am J Hum Genet 57:487–498

Thomson G, Robinson WP, Kuhner MK, Joe S, MacDonald MJ, Gottschall JL, Barbosa J, et al (1988) Genetic heterogeneity, modes of inheritance and risk estimates for a joint study of Caucasians with insulin dependent diabetes mellitus. Am J Hum Genet 43:799–816

Todd J (1995) Genetic analysis of type I diabetes using whole genome approaches. Proc Natl Acad Sci USA 92:8560–8565

Todd JA, Mijovic C, Fletcher J, Jenkins D, Bradwell AR, Barnett AH (1989) Identification of susceptibility loci for insulin dependent diabetes mellitus by trans-racial gene mapping. Nature 338:587–589

Valdes AM (1995) Population genetic modeling applied to human hereditary diseases. PhD thesis, University of California, Berkeley

Valdes AM, McWeeney S, Thomson G (1997) HLA class II DR-DQ amino acids and insulin-dependent diabetes mellitus: application of the haplotype method. Am J Hum Genet 59:717–728 (in this issue)

Vicario JL, Martinez-Laso J, Correll A (1992) Comparison between HLA-DRB and DQ DNA sequences and classical serological markers as type I (insulin dependent) diabetes mellitus predictive risk markers in the Spanish population. Diabetologia 35:475–481

Watanabe Y, Tokunaga K, Matsuki K, Takeuchi F, Matsuta K, Maeda H, Omoto K, et al (1989) Putative amino acid sequence of HLA-DRB chain contributing to rheumatoid arthritis susceptibility. J Exp Med 169:2263–2268

Yasunaga S, Kimura A, Hamaguchi K, Rønningen KS (1996) Different contribution of HLA-DR and -DQ genes in susceptibility and resistance to insulin dependent diabetes mellitus (IDDM). Tissue Antigens 47:37–48